

CLAIMS

1. An automated method for allocating resources among a plurality of resource-using computational entities in a data processing system, the method comprising:
establishing a service-level utility for each of said plurality of resource-using entities; and
transforming said service-level utility into a resource-level utility for each of said plurality of resource-using entities.
2. The method of claim 1, wherein the service-level utility is representative of an amount of business value obtained by each of said plurality of resource-using entities for various levels of performance and demand associated with the resource-using entity.
3. The method of claim 1, wherein the resource-level utility is representative of an amount of business value obtained by each of said plurality of resource-using entities when a quantity of said resources is allocated to the resource-using entity.
4. The method of claim 1, further comprising the steps of:
aggregating said resource-level utilities of all of said plurality of resource-using entities; and
computing a resource allocation from the aggregated utility information.
5. The method of claim 4, further comprising the step of:
executing and conveying to the plurality of resource-using entities said resource allocation.
6. The method of claim 1, wherein at least one of said plurality of resource-using entities operates to set its internal parameters, or an adjustable parameter of the resources the resource-using entity has been allocated so as to optimize the service-level utility, the resource-level utility, or both.

7. The method of claim 3, wherein the resource-level utility indicates, for at least one of said plurality of resource-using entities, a current utility based on current state descriptions of said at least one resource-using entity.
8. The method of claim 3, wherein the resource-level utility indicates, for at least one of said plurality of resource-using entities, an estimated cumulative discounted or undiscounted future utility starting from current state descriptions of said at least one resource-using entity.
9. The method of claim 8, wherein the estimated cumulative discounted or undiscounted future utility is based, for at least one of said plurality of resource-using entities, upon predictions of future state descriptions of said at least one resource-using entity.
10. The method of claim 8, wherein the estimated cumulative discounted or undiscounted future utility is trained on a temporal sequence of observed data using an adaptive machine learning procedure.
11. The method of claim 10, wherein the machine learning procedure is a reinforcement learning procedure.
12. The method of claim 11, wherein the reinforcement learning procedure is Q-Learning, Temporal Difference Learning, R-Learning or SARSA.
13. The method of claim 8, wherein the estimated cumulative discounted or undiscounted future utility is trained on a temporal sequence of observed data using a time-series prediction method.
14. The method of claim 4, wherein the step of aggregating said resource-level utilities of all of said plurality of resource-using entities is initiated by said plurality of resource-using entities.

15. The method of claim 4, wherein the step of aggregating said resource-level utilities of all of said plurality of resource-using entities is initiated by at least one resource arbiter adapted to compute said resource allocation from the aggregated utility information.

16. The method of claim 4, wherein the step of computing a resource allocation from the aggregated utility information comprises executing an optimization method to maximize a total utility of said data processing system.

17. The method of claim 16, wherein said optimization method comprises a standard linear or nonlinear algorithm.

18. The method of claim 17, wherein said optimization method is hill climbing, simulated annealing, linear programming or mixed-integer programming.

19. The method of claim 4, wherein the step of computing a resource allocation from the aggregated utility information comprises computing a cost that may be incurred in reallocating at least one of said resources from one of said plurality of resource-using entities to another.

20. The method of claim 1, wherein the resource-level utility is a function of client demand received by one of said plurality of resource-using entities and of a service-level agreement governing the performance of said one of said plurality of resource-using entities.

21. A computer readable medium containing an executable program for allocating resources among a plurality of resource-using computational entities in a data processing system, where the program performs the steps of:

establishing a service-level utility for each of said plurality of resource-using entities; and

transforming said service-level utility into a resource-level utility for each of said plurality of resource-using entities.

22. The computer readable medium of claim 21, wherein said program further performs the steps of:

aggregating said resource-level utilities of all of said plurality of resource-using entities; and

computing a resource allocation from the aggregated utility information.

23. The computer readable medium of claim 22, wherein said program further performs the step of:

executing and conveying to the plurality of resource-using entities said resource allocation.

24. The computer readable medium of claim 21, wherein at least one of said plurality of resource-using entities operates to set its internal parameters, or an adjustable parameter of the resources the resource-using entity has been allocated so as to optimize the service-level utility, the resource-level utility, or both.

25. The computer readable medium of claim 21, wherein the resource-level utility indicates an estimated cumulative discounted or undiscounted future utility starting from current state descriptions of said plurality of resource-using entities.

26. The computer readable medium of claim 21, wherein the resource-level utility indicates a current utility based on current state descriptions of said plurality of resource-using entities.

27. The computer readable medium of claim 21, wherein the resource-level utility indicates an estimated cumulative discounted or undiscounted future utility starting from current state descriptions of said plurality of resource-using entities.

28. The computer readable medium of claim 23, wherein the step of computing a resource allocation from the aggregated utility information comprises executing an optimization algorithm to maximize a business value of said data processing system.

29. A data processing system, comprising:
a plurality of entities adapted for processing client demands;
a plurality of resources adapted for allocation to said plurality of entities; and
at least one resource arbiter adapted for allocating said plurality of resources among said plurality of entities in a manner that optimizes a business value of the data processing system.
30. The data processing system of claim 25, wherein said plurality of entities are further adapted for transforming a respective service-level utility function into a corresponding service-level utility function.
31. The data processing system of claim 25, wherein said plurality of entities and said at least one resource arbiter are run on a single computer.
32. The data processing system of claim 25, wherein said plurality of entities and said at least one resource arbiter are run on different computers connected by a network.
33. The data processing system of claim 25, wherein said plurality of entities and said at least one resource arbiter are software modules comprising autonomic elements.
34. The data processing system of claim 25, wherein the data processing system is a server, a client computer or a network.